Final Project Report

Enzo Calcagno Daniel Gallant Bruno Marcoux

I. *Abstract*

Multilabel image classification is a growing area of interest in machine learning. In this final project, we consider the mechanism of class-specific residual attention (CSRA) proposed by Ke Zhu and Jianxin Wu of the University of Nanjing, as a computationally cheaper, simpler and more intuitive alternative to different attention based mechanisms. More specifically, we seek to verify some claims made by the authors, and explore variations. The main motivation of our experiment was to determine the effectiveness of CSRA on a different model than the ones tested in the paper. We chose VGG-16 as our backbone model and attempted to compare VGG-16 without CRSA to VGG-16 with CRSA by comparing their mAP scores. We found that the CSRA module proposed by the authors did lead to the expected performance on the given models.

II. *Introduction*

Breakthroughs in the world of image multilabel classification have relied on the adoption of RNN's, GCN's and Bayesian networks, all of which come at a great computational cost. The simplified CSRA module proposed by the authors consists of applying a combination of global average pooling with a global spatial max-pooling as the last step on a tensor resulting from an image that has been passed through a CNN backbone. The paper claims this results in class specific residual attention features which improve the accuracy of multilabel classification models. Equation 1 defines the CSRA feature f^i, for class i.

$$f^i = g + \lambda a^i \tag{1}$$

Where g, responsible for average pooling, is denoted the classical global class-agnostic feature, since its computation does not rely on any class:

$$g = \frac{1}{w\,x\,h} \sum_{k=1}^{w\,x\,h} x_k \tag{2}$$

Where, w and h refer to the height and width of tensor **x** which is shown in figure 1.

Symbol a^i, responsible for spatial pooling, refers to the class-specific feature vector for class i. It is the weighted combination of the tensor **x,** hence the attention in the module:

$$a^i = \sum_{k=1}^{w\,x\,h} s_k^i x_k \tag{3}$$

The coefficients s_k^i are class-specific attention scores for the i-th class at the j-th location and are defined using weights m, the tensor x, and temperature T in the following manner:

$$s_j^i = \frac{exp(T x_k^T m_i)}{\sum_{k=1}^{w\,x\,h} exp(T x_k^T m_i)} \tag{4}$$

In Equation 1, λ controls the amount of spatial pooling in CSRA. In its full version, as shown in figure 1, CSRA consists of H score tensors, and the residual attention operation is computed on each. The final logit vector $\hat{y}_0$ is obtained by summing all weighted CSRA features. Note that the authors define the sequence of temperatures to avoid making T a learnable parameter, which would complicate the process.

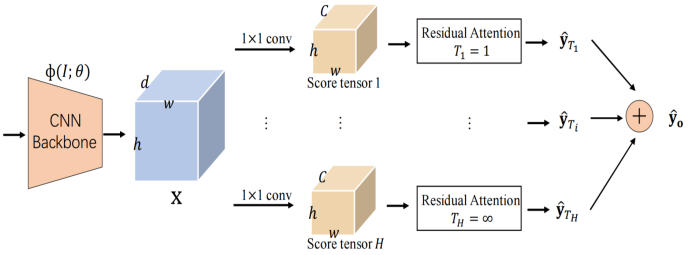

Figure 1. Full-Fledged CSRA module architecture

When a single head is chosen with the temperature parameter "T" tending to infinity, then the module simplifies to a simple modification which can be summarized to 4 lines of code, executed at testing time, resulting in virtually no extra computational cost. Note that in this case, spatial pooling becomes max pooling.

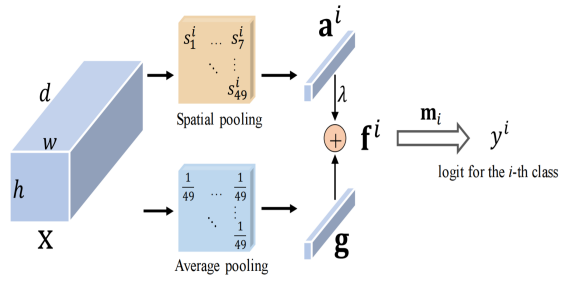

Figure 2. Simplified 1-headed CSRA module architecture.

The authors claim that these 4 lines of code consistently lead to improvement of multi-label recognition, across many diverse pretrained models and datasets, even without any extra training.

The authors claim that the a^i term in equation 1, global max pooling, which finds the maximum value among all spatial locations for each category, can be viewed as a class-specific attention mechanism. They conjecture it focuses the attention to classification scores at different locations for different object categories: which, if true, is ideal for multi-label recognition, especially when there are objects from many classes and/or with varying sizes.

VGG-16 is a popular classification and detection model and was chosen as the proposed new backbone to the CRSA model for our experiments.

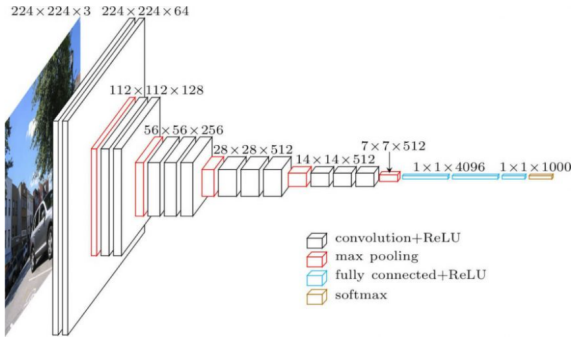

Figure 3. VGG-16 architecture

Our goal was to test whether we could improve the performance of VGG-16 on VOC2007 using the proposed CSRA model with VGG-16 as its backbone as is described in Figure 1.

III. Tasks

In this project we seek to verify some important results achieved by the authors. The specific tasks that were carried out for this paper are listed below:

Task 1: Use the 4 line fix the paper proposed on the provided datasets, validate we get the same results (this is done by setting H =1, and T = inf)

Task 2: Perform a control experiment to verify the individual impacts of average and global pooling by comparing $f^i = g$, $f^i = a^i$, and $f^i = g + \lambda a^i$.

Task 3: Perform hyperparameter tuning with the number of attention heads H to measure its impact on performance impacts.

Task 4: a) Use the VGG-16 pretrained model as a backbone to the CRSA model proposed in the paper and compare its mAP score to a VGG-16 model without CRSA.

IV. Results

In an attempt to maximize the variety of reproducibility testing, instead of reproducing all the tests provided in the report, for each task only one dataset-model combo was used.

Task 1: Validate accuracy results with CSRA

Dataset used: VOC2007.
Backbone model used: ResNet-cut, ResNet-10, VGG-16
These tests were conducted with the number of heads fixed to one and T set to infinity. The main mechanism the authors put forward.

Backbone	λ	mAP(%) No-CSRA	mAP(%) CSRA-original	mAP(%)CSRA-reproduced
ResNet -101	0.1	92.7 *(from list)*	94.7	94.6
ResNet -cut	0.1	93.9	95.2	95.1
VGG-16	1e-3	70.1	NA	NA

Table 1. Experimental results, task 1

Task 2: Control experiment
Dataset used: VOC2007.
Backbone model used: ResNet-cut

These tests were conducted with H=1 and T=1.

Method	Average Pooling	Spatial Max Pooling	mAP(%)
1	X		92.3
2		X	94.6
3	X	X	95.1

Table 2. Control experiment

Task 3: Hyperparameter tuning of H

Dataset used: VOC2007.
Backbone model used: ResNet-cut

Number of Heads	1	2	4	6	8
mAP(%)	95.1	95.2	95.2	95.1	95.2

Table 3. Hyperparameter tuning experiment

V. Discussion

1. Result Discussion

Task 1: For λ values of 0.1, for both of the explored CNN backbones we observe that our mAP's are 0.1% off the paper's reported mAP's. Additionally, in both cases, CSRA shows an improvement in mAP of 1.9% for first backbone and 1.2% for second backbone vs when no CSRA is applied.

Task 2: In order to validate the results found in the paper, a different dataset was used for the control experiment. While the authors used MS-COCO, we used VOC2007. The authors' conclusions were validated as we too found that using both average and max pooling led to the best results, as opposed to using to using any of the types of pooling in isolation

Task 3: As was the case in task 2, the VOC2007 dataset was used instead of the MS-COCO one. Once again, the results found by the authors were validated. However, we believe that correlation between number of heads and performance is too small to be considered significant. Unlike the authors, we do not share their opinion that this experiment is conclusive in confirming the link between number of heads and performance.

Task 4: Due to many unforeseen bugs and modifications needed to be done in the original project codebase, we were not able to run the VGG16_CSRA model as was described in Task 4, however we were able to get results for vgg-16 without the CRSA component as is shown in the results using a pretrained VGG-16 model trained on the VOC 2007 dataset.

2.Necessary details for reproducing the results, but were not specified in the original paper.

It was not specified in the original paper that the datasets used in the experiments must be downloaded and placed into a Dataset directory in the root of the project. Additionally, pretrained models must be individually downloaded with the appropriate code modifications in main.py, val.py and demo.py, however links are provided in the README file of the original projects' github repository

3.Summarize the key takeaways from the project and possibly directions for future investigation.

Overall for Resnet and Resnet cut we do still see an improvement in mAP score on the VOC2007 dataset in the reproducted experiments. We were able to obtain a VGG-16 score without CSRA to be used for comparison to a CSRA version of that model however we were unable to obtain data for the model. (figure 4)
For future investigation, the models proposed in the original paper should be tested on additional datasets, for instance the updated VOC 2012 dataset. The ideas proposed could moreover be applied to other popular multilabel classification models. Lastly, a different approach that we considered but did not go with would be to analyze the effect of modifying the relationship between the number of heads H and temperature T to see if the proposed CSRA model could be further improved upon. Also, another type of pooling could be used to complement the CSRA module, creating a new custom module.